Quantum Data Compression and Relative Entropy Revisited

Alexei Kaltchenko

Department of Physics and Computing, Wilfrid Laurier University, Waterloo, Ontario, N2L3C5, Canada.[*]

B. Schumacher and M. Westmoreland have established a quantum analog of a well-known classical information theory result on a role of relative entropy as a measure of non-optimality in (classical) data compression. In this paper, we provide an alternative, simple and constructive proof of this result by constructing quantum compression codes (schemes) from classical data compression codes. Moreover, as the quantum data compression/coding task can be effectively reduced to a (quasi-)classical one, we show that relevant results from classical information theory and data compression become applicable and therefore can be extended to the quantum domain.

I. INTRODUCTION

Quantum relative entropy, defined as $S(\rho||\sigma) \triangleq \mathrm{tr}(\rho \log \rho) - \mathrm{tr}(\rho \log \sigma)$, is always non-negative and is equal to zero if and only if $\rho = \sigma$, where ρ and σ are density operators defined on the same finite-dimensional Hilbert space. For a review of the properties of quantum relative entropy as well as the properties of quantum (von Neumann) entropy, defined as $S(\rho) = \mathrm{tr}(\rho \log \rho)$, see, for instance, [10].

Both quantum (von Neumann) entropy and quantum relative entropy can be viewed as quantum extensions of their respective classical counter-parts and inherit many of their properties and applications. Such inheritance however is often neither automatic, nor obvious. For example, classical information theory[4, Chap. 5] provides the following natural interpretations of Shannon entropy and relative entropy. A compression code (scheme), for example, Huffman code, which is made optimal for an information source with a given probability distribution p of characters, would require $H(p)$ bits per input letter to encode a sequence emitted by the source, where $H(p)$ stands for the Shannon entropy of the probability distribution p and is defined by

$$H(p) \triangleq -\sum p(x) \log p(x). \tag{1}$$

Thus, if using a compression code[1] (scheme) made optimal for an information source with probability distribution q, also called as "q-source", one could compress such a source with $H(q)$ bits per input character. If, to compress the same q-source, one uses a *different* compression code (scheme) optimal for a source with a *different* probability distribution p (i.e. p-source), it will require $H(q) + D(q||p)$ bits per input character, where $D(q||p)$ stands for the relative entropy function for the probability distribution q and p and is defined by

$$D(q||p) \triangleq \sum q(x) \log \frac{q(x)}{p(x)}. \tag{2}$$

Accordingly, the quantity $D(q||p)$ represents an additional encoding cost which arises from encoding a q-source with a code optimal for a p-source. This (classical) data compression interpretation of (classical) Shannon and relative entropies extends[8, 9] to the quantum domain if one replaces classical information sources having probability distributions p and q with quantum information sources having density matrices ρ and σ, respectively, as well as (classical) Shannon and relative entropies with (quantum) von Neumann and relative entropies, respectively.

In this paper, we provide an alternative, simple and constructive proof of this result. We build quantum compression codes (schemes) from classical compression codes. As the original source density matrix ρ can simply be replaced with $\tilde{\rho}$, a so-called "effective" source matrix[5] with respect to our computational basis, our quantum compression/coding task is effectively reduced to a (quasi-)classical one, with classical information sources having probability distributions $p_{\tilde{\rho}}$ and p_{σ}, which are formed by the eigenvalues of $\tilde{\rho}$ and σ, respectively. Thus, relevant results from classical information theory and data compression become applicable and therefore can be extended to the quantum domain.

---

[*]Electronic address: akaltchenko@wlu.ca
[1] See Subsection II A for the formal definitions of information sources, compression codes, etc.

Our paper is organized as follows. In Section II, we briefly review some classical data compression definitions and results, including information sources and source codes (also known as "data compression codes"). In Section III, we introduce quantum information sources, quantum data compression and discuss how classical data compression codes can be "turned" into quantum data compression codes. In Section IV, we introduce an "effective" density matrix for a quantum information source. In Section V, we state our main results: Lemma V.1 and Theorem 3 followed by a discussion. We conclude the paper with a brief conclusion section.

II. CLASSICAL DATA COMPRESSION

A. Classical Information Sources, Compression Codes, and Optimal Data Compression

Let $\mathcal{A} = \left\{ a_1, a_2, \ldots, a_{|\mathcal{A}|} \right\}$ be a data sequence alphabet, where the notation $|\mathcal{A}|$ stands for the cardinality of $\mathcal{A}$. Binary alphabet $\{0,1\}$ and Latin alphabet $\{a, b, c, d, \ldots, x, y, z\}$ are widely-used alphabet examples. We will also use the notation $|sequence\_name|$ to denote the sequence length. Let $\mathcal{A}^*$, $\mathcal{A}^+$, and $\mathcal{A}^\infty$ be respectively the set of all finite sequences, including the empty sequence, the set of all finite sequences of positive length, and the set of all infinite sequences, where all the sequences are drawn from $\mathcal{A}$. For any positive integer n, $\mathcal{A}^n$ denotes the set of all sequences of length n from $\mathcal{A}$. A sequence from $\mathcal{A}$ is sometimes called an $\mathcal{A}$-sequence. To save space, we denote an $\mathcal{A}$-sequence $x_1, x_2, \ldots, x_n$ by x^n for every integer $n > 0$.

Definition II.1 *An alphabet $\mathcal{A}$ independently and identically distributed (i.i.d.) information source is an i.i.d. discreet stochastic process, which is defined by a single-letter (marginal) probability distribution of characters from $\mathcal{A}$. We extend the single-letter probability distribution to a probability measure on the probability space constructed for the set $\mathcal{A}^+$ of all $\mathcal{A}$-sequences. We say "$\mathcal{A}$-source" or "p-source" to emphasize that the source has a certain alphabet $\mathcal{A}$ or a probability distribution p.*

Thus, we assume that every $\mathcal{A}$-sequence is generated by an i.i.d. information source.

Definition II.2 *An alphabet $\mathcal{A}$ source code $\mathbf{C}$ is a mapping from $\mathcal{A}^+$ into $\{0,1\}^+$, where, according to our notation, $\mathcal{A}^+$ is the set of all $\mathcal{A}$-strings and $\{0,1\}^+$ is the set of all binary strings.*

We restrict our attention to *uniquely decidable* codes that is, for every n and every $x_1, x_2, \ldots, x_n$, we have

$$\mathbf{C}^{-1}\left(\mathbf{C}(x_1, x_2, \ldots, x_n)\right) = x_1, x_2, \ldots, x_n$$

For every n and every $x_1, x_2, \ldots, x_n$, we call a sequence $\mathbf{C}(x_1, x_2, \ldots, x_n)$ a *codeword*. We often call a source code a *compression* code and use the terms "source code" and "compression code" interchangeably.

Definition II.3 *For any given source and an alphabet $\mathcal{A}$ source code $\mathbf{C}$, we define the compression rate of $\mathbf{C}$ in bits per input character by $\left\langle \frac{1}{n} |\mathbf{C}(x_1, x_2, \ldots, x_n)| \right\rangle_p$, where the length of a compressed sequence (codeword) $\mathbf{C}(x_1, x_2, \ldots, x_n)$ is divided by n and averaged with respect to information source's probability measure p.*

Theorem 1 (Optimal Data Compression) *For an i.i.d. information source with probability measure p and any compression code, the best achievable compression rate is given by Shannon entropy $H(p)$, which is defined by*

$$H(p) \triangleq -\sum_{x \in \mathcal{A}} p(x) \log p(x) \tag{3}$$

Definition II.4 *For any probability distribution p, we call a source code* optimal *for an i.i.d. source with probability distribution p if the compression rate for this code is equal to Shannon entropy $H(p)$. We denote such a code by $\mathbf{C}_p$.*

Remark 1 *Optimal compression code(s) exist for any and every source's probability distribution.*

B. Classical Data Compression and Relative Entropy

Suppose we have a compression code $\mathbf{C}_p$, that is $\mathbf{C}_p$ is optimal for an $\mathcal{A}$-source with a probability distribution p. What happens to the compression rate if we use $\mathbf{C}_p$ to compress another $\mathcal{A}$-source with a different distribution q?

It is not difficult to infer that the compression rate will generally increase, but by how much? The increase in the compression rate will be equal to relative entropy function $D(q||p)$, which is defined by

$$D(q||p) \triangleq \sum_{x \in \mathcal{A}} q(x) \log \frac{q(x)}{p(x)}, \tag{4}$$

where relative entropy function $D(q||p)$ is always non-negaitive[4] and equal to zero if and only if $q \equiv p$. More precisely, if using a code $\mathbf{C}_q$ (i.e. optimal for a distribution q), one could encode a q-source with $H(q)$ bits per input character. If, instead, one uses a code $\mathbf{C}_p$ (i.e. optimal for p) to encode a q-source, it will require $H(q) + D(q||p)$ bits per input character.

III.   TURNING CLASSICAL SOURCE CODES INTO QUANTUM TRANSFORMATIONS

In this section we review how a classical one-to-one mapping (code) can be naturally extended[2] to a unitary quantum transformation. But first, we need to define quantum information sources.

A.   Quantum Information Sources

Informally, a quantum identically and independently distributed (i.i.d.) information source is an infinite sequence of identically and independently generated quantum states (say, photons, where each photon can have one of two fixed polarizations). Each such state "lives" in its "own" copy of the same Hilbert space and is characterized by its "own" copy of the same density operator acting on the Hilbert space. We denote the Hilbert space and the density operator by $\mathcal{H}$ and ρ, respectively. Then, the sequence of n states "lives" in the space $\mathcal{H}^{\otimes n}$ and is characterized by the density operator $\rho^{\otimes n}$ acting on $\mathcal{H}^{\otimes n}$. For example, in case of a photon sequence, we have $\dim \mathcal{H} = 2$. Thus, to define a quantum source, we just need to specify ρ and $\mathcal{H}$. In this paper, we restrict our attention to quantum sources defined on two-dimensional Hilbert spaces. Often, just ρ is given and $\mathcal{H}$ is not explicitly specified as all Hilbert spaces of the same dimensionality are isomorphic.

B.   Turning Classical Source Codes into Quantum Transformations

Now we ready to explain how a classical source code can be turned into a quantum transformation. For the rest of the paper, we will only be considering *binary* classical sources with alphabet $\mathcal{A} = \{0, 1\}$ and quantum information sources defined on a 2-dimensional Hilbert space.

Let $\mathcal{B} \triangleq \{|0\rangle, |1\rangle\}$ be an arbitrary, but fixed orthonormal basis in the quantum information source space $\mathcal{H}$. We call $\mathcal{B}$ a *computational basis*. Then, for a sequence of n quantum states, the orthonormal basis will have 2^n vectors of the form

$$|e_1\rangle \otimes |e_2\rangle \otimes |e_3\rangle \otimes \cdots \otimes |e_n\rangle, \tag{5}$$

where, for every integer $1 \leqslant i \leqslant n$, $|e_i\rangle \in \mathcal{B} \triangleq \{|0\rangle, |1\rangle\}$. To emphasize the symbolic correspondence between binary strings of length n and the basis vectors in $\mathcal{H}^{\otimes n}$, we can rewrite (5) as

$$|e_1 e_2 e_3 \ldots e_n\rangle, \tag{6}$$

where, for every integer $1 \leqslant i \leqslant n$, $e_i \in \{0, 1\}$. In other words, if we add the Dirac's "ket" notation at the beginning and the end of any binary string of length n, then we will get a basis vector in $\mathcal{H}^{\otimes n}$.

Thus, for any computational basis $\mathcal{B} \in \mathcal{H}$, positive integer n, and a classical fully reversible function (mapping) for n-bit binary strings, we can define a unitary transformation on $\mathcal{H}^{\otimes n}$ as follows.

Let $\varphi : \{0, 1\}^n \to \{0, 1\}^n$ be a classical fully reversible function for n-bit binary strings. For every φ and n, we define a unitary operator $U_{\mathcal{B}, \varphi}^n : \mathcal{H}_{in} = \mathcal{H}^{\otimes n} \to \mathcal{H}_{out}$ by the bases vectors mapping $U_{\mathcal{B}, \varphi}^n |e_1 \ldots e_n\rangle = |\varphi(e_1 \ldots e_n)\rangle$, where, for every integer $1 \leqslant i \leqslant n$, $e_i \in \{0, 1\}$ and $|e_i\rangle \in \mathcal{B} \triangleq \{|0\rangle, |1\rangle\}$. We point out that $\mathcal{H}_{out}$ denotes an output space which is an isomorphic copy of the input space $\mathcal{H}_{in} = \mathcal{H}^{\otimes n}$, and $\mathcal{H}_{out}$ may sometimes coincide with $\mathcal{H}_{in}$.

---

[2] For more details, see [5] or [7]

An elegant way of designing quantum circuits to implement $U_{\mathcal{B},\varphi}^n$ was given by Langford in [7], where $U_{\mathcal{B},\varphi}^n$ was computed by the following two consecutive operations defined on $\mathcal{H}_{in} \otimes \mathcal{H}_{out}$:

$$|e_1 \ldots e_n\rangle|e_1^{out} \ldots e_n^{out}\rangle \longrightarrow |e_1 \ldots e_n\rangle|e_1^{out} \ldots e_n^{out} \oplus \varphi(e_1 \ldots e_n)\rangle \tag{7}$$

$$|e_1 \ldots e_n\rangle|e_1^{out} \ldots e_n^{out}\rangle \longrightarrow |e_1 \ldots e_n \oplus \varphi^{-1}(e_1^{out} \ldots e_n^{out})\rangle|e_1^{out} \ldots e_n^{out}\rangle, \tag{8}$$

where $e_i, e_i^{out} \in \{0,1\}$, $|e_1^{out} \ldots e_n^{out}\rangle \in \mathcal{H}_{out}$, and the notation $\oplus$ stands for bit-wise modulo-2 addition.

In the above, Relation (7) maps a basis vector $|e_1 \ldots e_n\rangle|\underbrace{0 \ldots 0}_{n}\rangle$ to the basis vector $|e_1 \ldots e_n\rangle|\varphi(e_1 \ldots e_n)\rangle$. Relation (8) maps a basis vector $|e_1 \ldots e_n\rangle|\varphi(e_1 \ldots e_n)\rangle$ to the basis vector $|\underbrace{0 \ldots 0}_{n}\rangle|\varphi(e_1 \ldots e_n)\rangle$.

Another interesting approach for computing $U_{\mathcal{B},\varphi}^n$ was suggested in [5], where Jozsa and Presnell used a so-called "digitalization procedure".

Remark 2 *We reiterate that, given a quantum source's Hilbert space $\mathcal{H}$, the quantum transformation $U_{\mathcal{B},\varphi}^n$ entirely depends on- and determined by the selection of the computational basis $\mathcal{B} \in \mathcal{H}$ and a classical fully reversible function φ (mapping) for n-bit binary strings. We also point out that as long as $\varphi(\cdot)$ and $\varphi^{-1}(\cdot)$ can be classically computed in linear time and space, then, $U_{\mathcal{B},\varphi}^n$ can be computed[1, 2, 3] in (log-)linear time and space, too.*

C. Quantum Data Compression Overview

We wish to compress a sequence of quantum states so that at a later stage the compressed sequence can be decompressed . We also want that the decompressed sequence will be the same as- or "close" to the original sequence. The measure of "closeness" is called *quantum fidelity*. There are a few different definitions of fidelity. One widely used is the fidelity between any pure state $|\psi\rangle$ and mixed state ρ, defined by

$$F(|\psi\rangle\langle\psi|, \rho) = \langle\psi|\rho|\psi\rangle$$

We now introduce a new notation $U_{\mathcal{B},\mathbf{C}}^n$ for a unitary transformation defined as $U_{\mathcal{B},\mathbf{C}}^n := U_{\mathcal{B},\varphi}^n\big|_{\varphi=\mathbf{C}}$. Thus, the unitary transformation $U_{\mathcal{B},\mathbf{C}}^n$ acts on a product space $\mathcal{H}^{\otimes n} \times \mathcal{H}^{\otimes n}$ and can be viewed as a quantum extension of the classical compression code $\mathbf{C}$. So to call $U_{\mathcal{B},\mathbf{C}}^n$, we will use terms "unitary transformation" and "(quantum) compression code" interchangeably.

Definition III.1 *Given a quantum i.i.d. source with density operator ρ, we define a compression rate of $U_{\mathcal{B},\mathbf{C}}^n$ by*

$$\left\langle \frac{1}{n} \left| U_{\mathcal{B},\mathbf{C}}^n |z_1 z_2 z_3 \ldots z_n\rangle \right| \right\rangle, \tag{9}$$

where $|z_1 z_2 z_3 \ldots z_n\rangle \in \mathcal{H}^{\otimes n}$ is a sequence emitted by the quantum source, and the length of compressed quantum sequence $U_{\mathcal{B},\mathbf{C}}^n |z_1 z_2 z_3 \ldots z_n\rangle$ is divided by n and averaged in the sense of a quantum-mechanical observable[3].

If we "cut" the compressed sequence $U_{\mathcal{B},\mathbf{C}}^n |z_1 z_2 z_3 \ldots z_n\rangle$ at the length "slightly more" than $\langle|U_{\mathcal{B},\mathbf{C}}^n |z_1 z_2 z_3 \ldots z_n\rangle|\rangle$, then it can be de-compressed with high fidelity. Thus, $\langle|U_{\mathcal{B},\mathbf{C}}^n |z_1 z_2 z_3 \ldots z_n\rangle|\rangle$ is the average number of qubits needed to faithfully represent an uncompressed sequence of length n. See [9] for a detailed treatment.

Theorem 2 (Schumacher's Optimal Quantum Data Compression[8]) *For a quantum i.i.d. source with a density operator ρ and any compression scheme, the best achievable compression rate with a high expected fidelity is given by the von Neumann entropy, which is defined by*

$$S(\rho) = tr(\rho \log_2 \rho) \tag{10}$$

Definition III.2 *A compression code, for which the above rate is achieved, is called* optimal.

Lemma III.1 *For a density operator ρ, we choose the eigenbasis of ρ to be our computational basis $\mathcal{B}$. Let $\mathbf{C}_{p_\rho}$ be a classical compression code, which is optimal for a probability distribution formed my the set of the eigenvalues of ρ. Then, a unitary transformation $U_{\mathcal{B},\mathbf{C}_{p_\rho}}^n$ is an optimal quantum compression code for a quantum information source with the density operator ρ.*

---

[3] Of course, the length a quantum sequence is a quantum-mechanical variable. See [9] for a detailed discussion on the subject

IV. "EFFECTIVE" SOURCE DENSITY MATRIX

We recall a quantum (binary) identically and independently distributed (i.i.d.) information source is defined by a density operator acting on a two-dimensional Hilbert space.

Let $\mathcal{B} = \{|b_i\rangle\}$ be our computational basis and suppose we have a quantum i.i.d. source with density operator ρ acting on our two-dimensional computational Hilbert space. Let $\sum_i \lambda_i |\lambda_i\rangle\langle\lambda_i|$ be an orthonormal decompositions of ρ.

For every ρ and $\{|b_i\rangle\}$, Jozsa and Presnell defined[5] a so-called "effective" source density matrix $\tilde{\rho}$ with respect to a (computational) orthonormal basis $\{|b_i\rangle\}$ as follows

$$\tilde{\rho} \triangleq \sum_j \eta_j |b_j\rangle\langle b_j|,$$

where $\eta_j \triangleq \sum_i P_{ij}\lambda_i$ and $[P_{ij}]$ is a doubly stochastic matrix defined by

$$P_{ij} = \langle\lambda_i|b_j\rangle \, \langle b_j|\lambda_i\rangle \geqslant 0.$$

We observe the following obvious fact. If the source eigenbasis coincides with the computation basis, i.e. $\{|\lambda_i\rangle\} = \{|b_i\rangle\}$, then effective source density matrix coincides with the actual source density matrix, i.e. $\tilde{\rho} = \rho$.

V. MAIN RESULT

Definition V.1 *With any density matrix ρ, we associate a probability distribution formed by the density matrix's eigenvalues and denote it by putting the density matrix notation as a subscript: p_ρ. That is $p_\rho = \{\lambda_0, \lambda_1\}$, where $\{\lambda_0, \lambda_1\}$ are the eigenvalues of ρ.*

Lemma V.1 *Let ρ and σ be density matrices defined on the same Hilbert space, with orthonormal decompositions $\rho = \sum_i \lambda_i \, |\lambda_i\rangle\langle\lambda_i|$ and $\sigma = \sum_j \chi_j \, |\chi_j\rangle\langle\chi_j|$.*
Let $\tilde{\rho}$ be the effective density matrix of ρ with respect to the basis $\{|\chi_i\rangle\}$, and let $\{\eta_j\}$ be the eigenvalues of $\tilde{\rho}$.
Then, we have the following relation between quantum and classical entropies:

$$S(\rho||\sigma) + S(\rho) = D(p_{\tilde{\rho}}||p_\sigma) + H(p_{\tilde{\rho}}), \tag{11}$$

where $D(\cdot||\cdot)$ is classical relative entropy, $H(\cdot)$ is classical (Shannon) entropy, $p_{\tilde{\rho}}$ and p_σ are the probability distributions formed by eigenvalues sets of $\tilde{\rho}$ and σ, respectively. That is we have $p_{\tilde{\rho}} = \{\eta_j\}$ and $p_\sigma = \{\chi_i\}$.

Proof 1 *By the definition of effective density matrix, we have*

$$\eta_j \triangleq \sum_i P_{ij}\lambda_i,$$

where $[P_{ij}]$ is a doubly stochastic matrix, defined by

$$P_{ij} = \langle\lambda_i|\chi_j\rangle \, \langle\chi_j|\lambda_i\rangle.$$

By the definition of quantum relative entropy, we have

$$S(\rho||\sigma) \triangleq tr(\rho \log \rho) - tr(\rho \log \sigma).$$

First, we look at the quantity $tr(\rho \log \sigma)$:

$$tr(\rho \log \sigma) = tr(\rho \log \sigma \sum_i |\lambda_i\rangle\langle\lambda_i|) = \sum_i tr(\rho \log \sigma |\lambda_i\rangle\langle\lambda_i|) = \sum_i \langle\lambda_i|\rho \log \sigma|\lambda_i\rangle.$$

Substituting the identity $\langle\lambda_i|\rho = \lambda_i\langle\lambda_i|$ into the above equation, we get

$$tr(\rho \log \sigma) = \sum_i \lambda_i \, \langle\lambda_i| \log \sigma|\lambda_i\rangle \tag{12}$$

On the other hand, we have the following identity for the log function:

$$\log \sigma = \sum_j \log(\chi_j) \, |\chi_j\rangle\langle\chi_j| \tag{13}$$

and, therefore, substituting the right-hand-side of (13) into $\langle\lambda_i| \log \sigma |\lambda_i\rangle$, we get

$$\langle\lambda_i| \log \sigma |\lambda_i\rangle = \langle\lambda_i| \left(\sum_j \log(\chi_j) \, |\chi_j\rangle\langle\chi_j| \right) |\lambda_i\rangle = \sum_j \log(\chi_j) \, \langle\lambda_i|\chi_j\rangle \, \langle\chi_j|\lambda_i\rangle = \sum_j \log(\chi_j) \, P_{ij} \tag{14}$$

Combining equations (12) and (14), we obtain

$$tr(\rho \log \sigma) = \sum_i \lambda_i \sum_j \log(\chi_j) \, P_{ij} = \sum_j \log(\chi_j) \sum_i P_{ij}\lambda_i = \sum_j \eta_j \, \log(\chi_j). \tag{15}$$

Substituting the right-hand-side of the quantum entropy identity $S(\rho||\sigma) = \sum_i \lambda_i \log \lambda_i$ and that of (15) into the definition of quantum relative entropy, we have

$$S(\rho||\sigma) = \sum_i \lambda_i \log \lambda_i - \sum_j \eta_j \, \log \chi_j = -S(\rho) - \sum_j \eta_j \, \log \chi_j = -S(\rho) - \sum_j \eta_j \, \log \chi_j + \sum_j \eta_j \, \log \eta_j - \sum_j \eta_j \, \log \eta_j \tag{16}$$

Thus, we get

$$S(\rho||\sigma) = -S(\rho) + D(\eta_j||\chi_j) + H(\eta_j), \tag{17}$$

where $D(\eta_j||\chi_j)$ is the (classical) relative entropy of the probability distributions η_j and χ_j; $H(\eta_j)$ is the (classical) entropy of the distribution η_j. We can rewrite (17) as (11) which completes the proof.

In Subsection II A, we have introduced a notation $\mathbf{C}_p$ to denote a classical data compression code (mapping) optimal for a source probability distribution p. Then, according to our notation, the code $\mathbf{C}_{p_\sigma}$ will be optimal for a (classical) i.i.d. source with probability distribution formed by the eigenvalues of a density matrix σ.

Lemma V.2 (Classical data compression and relative entropy) *Let $z^n \triangleq z_1, \ldots, z_n$ be a sequence emitted by a classical i.i.d. source with a marginal probability distribution formed by the eigenvalues of $\tilde{\rho}$. Then, from classical data compression theory, we have*

$$\left\langle \frac{1}{n} \, |\mathbf{C}_{p_\sigma}(z^n)| \right\rangle_{p_{\tilde{\rho}}} = D(p_{\tilde{\rho}}||p_\sigma) + H(p_{\tilde{\rho}}), \tag{18}$$

where the notation in the left-hand-side stands for the length of compressed sequence (codeword) divided by n and averaged with respect to the probability measure $p_{\tilde{\rho}}$.

Combining Lemma V.1, Lemma V.2, and the results of Section III, we obtain the following theorem:

Theorem 3 (Quantum Data Compression and Relative Entropy) *Let $|z_1 z_2 z_3 \ldots z_n\rangle \in \mathcal{H}^{\otimes n}$ be emitted by a quantum i.i.d. source with density operator ρ defined on a Hilbert space $\mathcal{H}$. Let σ be an arbitrary density operator defined on the same Hilbert space $\mathcal{H}$.*

We choose the eigenbasis of σ to be our computational basis $\mathcal{B}$, and let $\tilde{\rho}$ be the effective matrix of ρ with respect to $\mathcal{B}$. Then, the following relations hold:

$$S(\rho||\sigma) + S(\rho) = D(p_{\tilde{\rho}}||p_\sigma) + H(p_{\tilde{\rho}}) = \left\langle \frac{1}{n} \left| U^n_{\mathcal{B}, \mathbf{C}_{p_\sigma}} \, |z_1 z_2 z_3 \ldots z_n\rangle \right| \right\rangle, \tag{19}$$

where the notation in the right-hand-side stands for the length of compressed quantum sequence divided by n and averaged in the sense of a quantum-mechanical observable.

Remark 3 *Quantum relative entropy $S(\rho||\sigma)$ has the following operational meaning. If a quantum compression code, optimal for one quantum information source with a density operator σ, is applied to another source with a density operator ρ, then the increase of the compression rate is equal to $S(\rho||\sigma)$.*

Remark 4 *By fixing ρ and the computational basis $\mathcal{B}$ (i.e. the eigenbasis of σ), we will also fix $\tilde{\rho}$. Then, for a fixed $\tilde{\rho}$, $D(p_{\tilde{\rho}}||p_\sigma)$ is minimized (and is therefore equal to zero) when $\sigma = \tilde{\rho}$. Then, as it follows from Theorem 3, for an arbitrary computational basis $\mathcal{B}$, the best possible compression rate is equal to $H(p_{\tilde{\rho}}) = S(\tilde{\rho})$ and is achieved with transformation $U^n_{\mathcal{B}, \mathbf{C}_{\tilde{\rho}}}$. In [5], such setup is called "mismatched bases" compression.*

If we choose the eigenbasis of ρ to be our computational basis $\mathcal{B}$, then we have $\tilde{\rho} = \rho$, and optimal quantum compression is achieved with $U^n_{\mathcal{B}, \mathbf{C}_{p_\rho}}$, with compression rate $H(p_\rho) = S(\rho)$.

Discussion 1 *As long as $U^n_{\mathcal{B}, \mathbf{C}_{p_\sigma}}$ is computed in a computational basis $\mathcal{B}$, we can simply replace the original source density matrix ρ with $\tilde{\rho}$, the effective one with respect to the computational basis. Then, our quantum compression/coding task is effectively reduced to a classical one, with classical information sources having probability distributions $p_{\tilde{\rho}}$ and p_σ. Therefore, all relevant results from classical information theory and data compression become applicable. We have already seen one such result, the statement of data compression (non-)optimality $D(p_{\tilde{\rho}}||p_\sigma) + H(p_{\tilde{\rho}})$.*

To give another example, we consider the following task of quantum relative entropy estimation. Suppose we have two quantum i.i.d. sources with unknown density matrices ρ and σ, where ρ and σ defined on the same Hilbert space. We want to estimate quantum relative entropy $S(\rho||\sigma)$.

Let $|z_1 z_2 z_3 \ldots z_n\rangle$ and $|x_1 x_2 x_3 \ldots x_n\rangle$ be emitted by the sources with density matrices ρ and σ, respectively. From Theorem 3, one could (hope) to estimate the quantity $S(\rho||\sigma) + S(\rho)$ by simply measuring the average length of codeword $U^n_{\mathcal{B}, \mathbf{C}_{p_\sigma}} |z_1 z_2 z_3 \ldots z_n\rangle$, then estimate $S(\rho)$ by measuring the average length of codeword $U^n_{\mathcal{B}, \mathbf{C}_{p_\rho}} |z_1 z_2 z_3 \ldots z_n\rangle$, and then subtract the latter from the former. For such straightforward approach to work, one should have quantum compression codes $U^n_{\mathcal{B}, \mathbf{C}_{p_\sigma}}$ and $U^n_{\mathcal{B}, \mathbf{C}_{p_\rho}}$, which are optimal for sources with density matrices ρ and σ, respectively. But as ρ and σ are unknown, so are $U^n_{\mathcal{B}, \mathbf{C}_{p_\sigma}}$ and $U^n_{\mathcal{B}, \mathbf{C}_{p_\rho}}$. Fortunately, in classical data compression, there exist so-called "universal compression codes"[11], which will let estimate (classical) quantities $D(p_{\tilde{\rho}}||p_\sigma) + H(p_{\tilde{\rho}})$ and $H(p_{\tilde{\rho}})$ even without knowing probability distributions $p_{\tilde{\rho}}$ and p_σ. So, based on these classical codes, we can construct quantum codes[6] (unitary transformations) to estimate $S(\rho||\sigma)$.

VI. CONCLUSION

We have provided a simple and constructive proof of a quantum analog of a well-known classical information theory result on a role of relative entropy as a measure of non-optimality in data compression. We have constructed quantum compression codes (schemes) from classical data compression codes. Moreover, as the quantum data compression/coding task can be effectively reduced to a (quasi-)classical one, we show that relevant results from classical information theory and data compression become applicable and therefore can be extended to the quantum domain.

[1] C. Bennett, "Logical reversibility of computation", *IBM J. Res. Devel.* Vol. 17, p. 525, 1973.

[2] E. Bernstein and U. Vazirani, "Quantum complexity theory", *Special issue on Quantum Computation of the Siam Journal of Computing*, Oct. 1997

[3] H. Buhrman, J. Tromp, and P. Vitanyi, "Time and Space Bounds for Reversible Simulation", *Proc. International Conference on Automata, Languages and Programming*, Lecture Notes in Computer Science, Springer-Verlag, Berlin, 2001.

[4] T. Cover, J. Thomas, *Elements of information theory*, Wiley-Interscience, New York, 1991.

[5] R. Jozsa and S. Presnell, "Universal quantum information compression and degrees of prior knowledge", *Proc. Roy. Soc. London Ser. A*, Vol. 459, pp. 3061–3077, 2003.

[6] A. Kaltchenko, "Universal Non-destructive Estimation of Quantum Relative Entropy", *Proc. of the 2004 IEEE International Symposium on Information Theory*, p. 356, USA, 2004.

[7] J. Langford, "Generic quantum block compression", *Physical Review A*, Vol. 65, p. 52312, 2002.

[8] B. Schumacher, "Quantum Coding", *Physical Review A*, Vol. 51, p. 2738, 1995.

[9] B. Schumacher and M.D. Westmoreland, "Indeterminate length quantum coding", *Physical Review A*, Vol. 64, p. 42304, 2001.

[10] V. Vedral, "The role of relative entropy in quantum information theory", *Reviews of Modern Physics*, Vol. 74, p. 197, 2002.

[11] J. Ziv and N. Merhav, "A measure of relative entropy between individual sequences with application to universal classification", *IEEE Trans. Inform. Theory*, Vol. 39, No. 4, pp. 1270–1279, July 1993.